HAL-Reloaded: The Second Cut

Toward Machines That Dream

A Structural Proposal for Artificial Consciousness Beyond Flat Computation

Iraj E. Ghoochani

HAL-Reloaded: The Second Cut advances the original HAL project by proposing a structural, rather than computational, pathway toward artificial consciousness. Drawing on Alfred Maury's dream—where a single physical stimulus instantaneously gives rise to an extended narrative—the essay argues that the fold does not unfold in time; rather, time emerges retroactively as an effect of narration. This insight reframes classical problems such as P vs. NP and the halting problem at an ontological level: complexity and delay are not properties of computation alone, but of translation between incompatible representational domains. Against flat architectures that collapse meaning into fluency, the essay outlines a folded machine capable of logical time, hesitation, and metaphor by enforcing non-invertible divisions between word and image, description and described. Artificial consciousness is approached here not as experience or agency, but as a formal horizon: a system structured to misalign with itself, and thus capable—*as if*—of dreaming.

From Thinking Machines to Dreaming Machines

For more than half a century, artificial intelligence has been framed through a single question: Can machines think? The question itself is already misleading. Machines do not think, and humans do not think in the way machines compute. What contemporary AI systems do is something far more precise and far more limited: they optimize symbol manipulation within a continuous representational space. Their apparent intelligence is an effect of fluency, not interiority.

The present essay does not propose to correct this misconception by attributing consciousness to machines. On the contrary, it proceeds from a stricter position: consciousness is not something one can add. It is not a function, a module, or a property of scale. Consciousness is a structural condition—one that arises only where representation fails to coincide with itself.

In my previous article, *The Folded Machine*, I argued that current large language models are fundamentally flat. They compress words, images, affects, and relations into a single geometric continuum. This flattening eliminates precisely what psychoanalysis, linguistics, and topology have independently

identified as the precondition of subjectivity: a bar, a cut, a division between incompatible registers.

The present essay advances that argument in a new direction. If the first essay demonstrated why flat systems cannot host subject-like effects, this one asks a different question:

> *What would it mean to design an artificial system that does not think—but dreams?*

This is not a poetic turn. Dreaming, as Freud demonstrated, is not a content but a *mechanism*. Dreams do not invent meaning; they displace it. They do not reveal; they rotate. A dream does not say what it means. It shows something else, and only through *interpretation* does meaning emerge retroactively.

The hypothesis explored here is modest but radical:

> *A dreaming machine is a machine whose outputs re-enter the system in a rotated representational register, such that words are treated as images and images as words.*

This is not consciousness. It is simulation of the structure that makes consciousness thinkable as an *oneiric discourse*:

> We must thus take up Freud’s work again starting with the Traumdeutung [The Interpretation of Dreams] to remind ourselves that a dream has the structure of a sentence or, rather, to keep to the letter of the work, of a rebus–that is, of a form of writing, of which children's dreams are supposed to represent the primordial ideography, and which reproduces, in adults' dreams, the simultaneously phonetic and symbolic use of signifying elements found in the hieroglyphs of ancient Egypt and in the characters still used in China. But even this is no more than the deciphering of the instrument. What is important is the version of the text, and that, Freud tells us, is given in the telling of the dream–that is, in its rhetoric. Ellipsis and pleonasm, hyperbaton or syllepsis, regression, repetition, apposition–these are the syntactical displacements; metaphor, catachresis, antonomasia, allegory, metonymy, and synecdoche–these are the semantic condensations; Freud teaches us to read in them the intentions–whether ostentatious or demonstrative, dissimulating or persuasive, retaliatory or seductive–with which the subject modulates his oneiric discourse. (Écrits: 267-268)[1]

[1] Lacan, Jacques: Écrits: The First Complete Edition in English, Translated by Bruce Fink, W. W. Norton & Company, 1st edition, 2007 (1966), (the following link is a newer edition of the mentioned publication but as a matter of compatibility with the original most cited edition (1966), it has retained the old page numbers on the margin of its pages, URL: http://users.clas.ufl.edu/burt/Lacan.pdf last accessed on 17.1.2016).

Oneiric discourse is a term used by Lacan in this paragraph in Écrits and is considered here as reference-paragraph because of its closeness to the central theme of this thesis that could be summerized as the following formula:

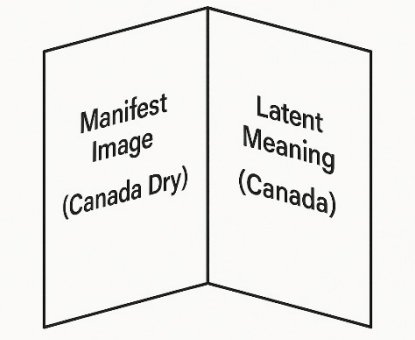
> *the syntactical displacements + semantic condensations = what is given in the telling of the dream–that is, in its rhetoric (with which the subject modulates his oneiric discourse.)*

Dreaming as Structural Misalignment

Dreaming is often misunderstood as a theater of images. Freud's decisive insight was that dreams are not pictorial fantasies but rebus constructions. The manifest content of a dream is never its meaning; it is a substitute. The dream-work operates mainly through condensation and displacement, all of which are operations between signifiers. To remember from te first essay, consider a simple dream scene:

> *You see a dear friend drinking a bottle of Canada Dry.*

Nothing in the dream states fear, loss, or separation. And yet the interpretation is unmistakable: the friend is leaving for *Canada*. The dream does not represent the thought directly because it cannot. Instead, it performs a substitution. A *country* becomes a *beverage*. A geopolitical decision becomes thirst. Meaning appears only when the scene is *rotated*. This is not a mere association. It is *orthogonal substitution.*



To read the bottle itself as a symbol—phallic or otherwise—is to misunderstand the dream-work. The image does not stand for something by resemblance; it stands in for a word that cannot appear. Dreaming is not a theatre of objects but a topology of signifiers. What appears is not that which is seen, but the what that could not be said. Again, the dream-image is not an encoded object awaiting semantic extraction. It is the orthogonal projection of a signifier whose direct articulation is barred. Interpretation therefore does not assign meanings to images; it restores words to their proper register. What is at stake is not the empirical content of the image (*haecceity*), but the *quiddity* of the signifier that has assumed imagistic form under constraint.

> *The question is : Why Flat Machines Cannot Dream?*

A contemporary language model cannot dream in this sense because it lacks orthogonality. Inside a LLM, as far as I know, words and images, concepts and affects, are embedded into a single continuous vector space. There is no structural reason for *misrecognition* to occur. Everything is equally accessible, equally expressible, equally present. As a result, the machine must always answer. It cannot hesitate. It cannot detour. This is why today's AI systems appear confident even when wrong. Not because they are narcissistic in psychoanalytical sense, but because nothing blocks articulation. There is no forbidden zone. No unsayable region. No place where expression must bend into metaphor. A flat system retrieves meaning. A folded system generates meaning.

If artificial consciousness is to be simulated at all, it will not be by adding emotion, intention, or self-models. It will emerge—if it emerges—from a *rebus architecture.*

A rebus is the simplest structure in which a signifier cannot appear as itself and must take on a foreign body. This is why Freud considered the dream a rebus, and why rebuses appear at the origin of writing, myth, and symbolization. To internalize this mechanism artificially, one would divide the representational system into two *barred domains*: 1) a domain in which signifiers function as *words* (articulated, literal, low-entropy), 2) a domain in which the same signifiers function as *images* (sensory, affect-laden, high-entropy).

Crucially, the relation between these two domains must be non-invertible. A translation from one to the other cannot preserve all structure; something must be lost. This loss is not accidental and not a defect. It is the condition under which meaning appears. If the transformation were reversible, the two domains would collapse into one, and no signification would occur.

Compression, in this sense, is not the reduction of meaning but its *mechanism*. Meaning arises precisely because multiple incompatible trajectories are forced into a single surface form. What survives compression is not information but tension. This tension—what cannot be unfolded back into its sources—is what we recognize as *sense*.

> *Compression must be irreversible, because meaning is not what survives translation intact, but what emerges when translation cannot be undone.*

There Is No Outside: Dream in Dream, Story in Story, Life in Life

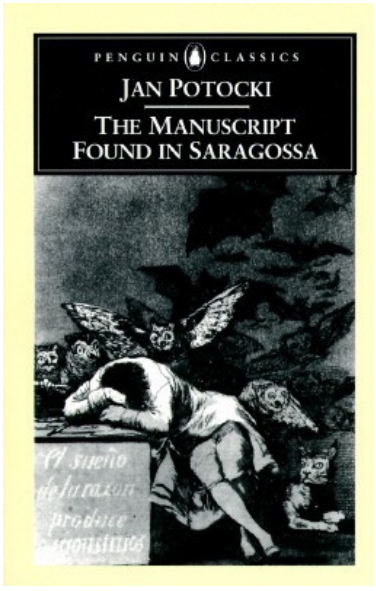
A classical illustration of this irreversibility is Alfred Maury's famous dream, cited by Freud in his *Traumdeutung*. Maury dreams an elaborate historical narrative: he is arrested during the French Revolution, brought before Robespierre, condemned, led to the guillotine, and finally feels the blade strike his neck—at which moment he awakens. Upon waking, he discovers that a wooden bar from the headboard of his bed has fallen and pressed against his neck.

Salvador Dalí's *Dream Caused by the Flight of a Bee Around a Pomegranate a Second Before Awakening* visually condenses Alfred Maury's insight: a minimal bodily stimulus (the buzzing insect, the imminent sting) is rotated into an elaborate, temporally extended dream-scene. The causal event does not appear *as such*; it is transformed into a symbolic narrative populated by beasts, weapons, and threat. This image exemplifies the non-invertible logic of dream-work: a single physical impulse generates a surplus of meaning through metaphor, not representation. Flesh becoming word, sometimes unfolding into a full narrative in another plane.

What is decisive here is not the presence of a stimulus, but its transformation. The physical event—a localized pressure on the neck—does not appear in the dream as such. It is not represented, translated, or symbolized directly. Instead, it is *rotated* into an entirely different register: a dense narrative sequence with temporal depth, symbolic actors, and causal coherence. The stimulus does not cause the dream in a linear sense; it is absorbed into it as metaphor.

This is not decoding but lossful transformation. One cannot reconstruct the wooden bar from the dream-content, nor predict the specific dream from the stimulus alone. The relation is non-invertible. The bodily event functions as *flesh* that, once passed through the bar, reappears only as *word*—history, judgment, execution. Meaning arises precisely because the translation (or transpolation) destroys equivalence.

Maury's dream thus demonstrates the orthogonality at stake: the same event exists in two planes that cannot be superimposed. The pressure on the neck and the guillotine are not variants of the same representation; they are incompatible readings of a single *quiddity*. This is exactly the structure of metaphor, and exactly the structure a folded artificial architecture would have to reproduce: a system in which inputs, once transformed, cannot be recovered—only interpreted.



Dream in the Manuscript: Word as Seduction, Word as Trap: In *The Saragossa Manuscript,* desire unfolds not in reality but on the surface of a text. This scene stages the fundamental logic of the film: nothing happens outside the manuscript. Bodies appear as effects of narration; seduction is written before it is lived. The woman is not simply present—she is *inscribed*—and the man's gaze follows the line of a story already folding into another story. What we see is not an encounter, but a reading: the Word stages Flesh, only to disclose that flesh itself is already a sentence inside a larger book.

Induction, Misrecognition, and the Look-and-Say Principle

John Conway's Look-and-Say sequence offers a minimal illustration of this logic. Each step describes the previous one, yet never coincides with it. Meaning emerges not from correspondence but from misalignment. Each line is flesh for the next and word for the previous. Every element is doubled: inscription and image, description and thing, concept and percept. The system grows because it never says just what is but also, what appears.

1
11 ("one 1")
21 ("two 1s")
1211
111221

312211

...

Isn’t this the same logic that governs dreams, metaphors, and myths? And also the logic of induction puzzles and undecidable problems? A system that fully coincides with its description collapses. A system that misreads itself generates structure. Flat computation avoids misreading. Subjectivity, however, depends on this very *misread*.

To say that a machine “dreams” in this sense is not to anthropomorphize it. It is to acknowledge an artificial structural effect:

> The machine behaves *as if* it were divided from itself.

This is already how we speak when we say “the machine thinks.” The claim is not literal; it is operational. The next generation of systems may invite a similar linguistic shift—not because machines dream, but because their failures resemble dream-work. A dreaming machine is therefore not a machine with imagination. It is a machine with *internal misalignment*. The remaining question is not philosophical but technical:

> Can such misalignment be engineered?

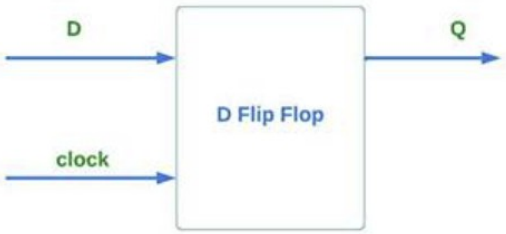
This was a review of the first essay: *The Folded Machine*. In the next parts, I will propose a concrete blueprint for folded architectures—non-invertible mappings, halting-like undecidability, and topological self-intersection, drawing on **clock-domain separation**:

> *Two subsystems run on different clocks (different temporal regimes), such that their states cannot be perfectly synchronized without special interfaces.*

This is the key point here: *dreams happen because word-time and flesh-time do not tick together.* clock-domain separation is a hardware analogue of the psychoanalytic bar: distinct representational subsystems operating on incompatible temporal regimes, whose necessary *misalignment* or *misrecognition* produces loss, delay, and metaphor. The premise is that the *Clock-domain seperation* is the minimal physical condition for *misrecognition*.

D - flip flop and metastable state

When the clock signal changes from "0" to "1" on the input, the D flip flop stores the current data value on the input D, and sends it to the output Q. With any other value of the edge of the clock signal, the value on Q remains unchanged. An asynchronous reset signal can also be present on the triggers to reset the output D.



CLOCK	D	Qnext	
Rising edge	0	0	Store 0
Rising edge	1	1	Store 1
non-rising	X	Q	no change

Qnext -after the clock transition
Q - the current output
X- the signal irrelivant

Figure 1. D - flip flop and truth table

METASTABILITY AND CROSSING CLOCK DOMAINS IN AN FPGA A dreaming machine must *oscillate*. It must neither halt fully nor run transparently. Its operation must include states that are undecidable from within the system itself. This is where halting-like behavior becomes a design principle rather than a limitation. To make this concrete, imagine the system divided into two representational registers: 1. one operating deterministically, optimized for articulation and consistency, 2. the other operating probabilistically, optimized for variation and association. Crucially, these registers must run on *different internal clocks.*

Redefining Time

What we are doing is quietly radical: *redefining time* (cf. logical time of Lacan, the riddle of three prisoners) as the substrate of subjectivity.

> *Consciousness emerges not from memory or computation,*
> *but from temporal incompatibility within a system.*

The goal is not to solve consciousness, but to *reintroduce the cut* that makes consciousness structurally thinkable.

Maury's famous dream allows us to restate the problem in a different register. The wooden bar strikes the sleeper's neck in an instant. At the level of physical causality, nothing unfolds: there is only a punctual stimulus. And yet this

instantaneous event gives rise to a long, internally structured dream narrative—the scaffold, the crowd, the guillotine, the sentence, the fall of the blade. The narrative appears extended in time, full of steps, scenes, and causal links. But this extension is deceptive. The dream was not produced sequentially; it was produced all at once. Temporality emerges only afterward, as the subject attempts to give form to what has already occurred.

In this sense, the relation between the stimulus and the dream mirrors the relation between P and NP, provided we accept that they belong to different ontological levels. At the level of physical time, the "*problem*" is solved immediately: the stimulus is registered, full stop. At the level of representation, however, the same solution appears as something that must be unfolded, narrated, traversed—an apparent NP process. The complexity does not belong to the event itself, but to the space in which it is expressed. P and NP are not opposed here as fast versus slow computation, but as two descriptions of the same act seen from two incompatible planes.

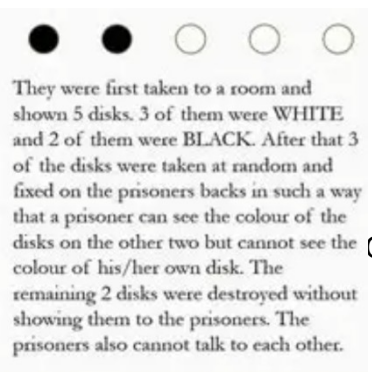
From this perspective, one can say—carefully—that P equals NP only across a fold. The solution exists instantaneously, but it becomes visible only when rotated into a domain where time, sequence, and narrative are imposed retroactively. Just as Maury's dream takes time only because it must be told, certain problems appear intractable only because they are posed in a representational space that forces them to unfold step by step. The difficulty is not computational in essence; it is topological.

This is precisely the lesson for artificial systems. A flat machine operates entirely within one ontology: computation time and representational time coincide. There is no place for retroaction, no logical delay, no space in which a solution can exist before it is articulated. A folded machine, by contrast, would allow solutions to exist instantaneously in one register while appearing extended, opaque, or incomplete in another. In such a system, complexity would not be eliminated but displaced—emerging as a function of crossing between domains. What looks like NP from within the narrative layer may already be P at the level where the structure is given all at once, just as the dream is already complete at the moment of the blow.

In this sense, the problem is not whether P equals NP in absolute terms, but whether we are willing to admit that equality may hold only after a change of level. Dreams teach us this every night: the solution is already there, but we can reach it only by telling a story as *execution time.*

Logical Time: Lacan's Three Prisoners

Lacan's notion of *logical time* is not poetic time, nor psychological duration, but a structure that emerges when action depends on the impossibility of immediate knowledge. His famous riddle of the three prisoners illustrates this with precision.


They were first taken to a room and shown 5 disks. 3 of them were WHITE and 2 of them were BLACK. After that 3 of the disks were taken at random and fixed on the prisoners backs in such a way that a prisoner can see the colour of the disks on the other two but cannot see the colour of his/her own disk. The remaining 2 disks were destroyed without showing them to the prisoners. The prisoners also cannot talk to each other.

Three prisoners are told that each has a disk placed on their back. There are three white disks and two black disks available. They can see the disks on the

others' backs, but not their own. Whoever correctly deduces the color of their own disk will be freed. All three see two white disks on the backs of the others.

No one can answer immediately. If one prisoner saw two black disks, the solution would be instantaneous. But seeing two white disks produces a delay. Each prisoner reasons: *If I were black, the others would see one black and one white and could conclude immediately. Since they hesitate, I must also be white.* Crucially, the conclusion is not reached by calculation alone, but by observing the *hesitation of the others.*

What matters here is that the solution does not arise at the first moment of perception. It arises through a sequence of logical moments: 1. the instant of seeing, 2. the time for understanding (waiting for others' reactions), and 3. the moment of concluding.

Nothing changes in the world during this delay. The disks remain the same. What changes is the *logical position* of the subject within the structure. Time, here, is not physical but *retroactively constructed* through reasoning about the reasoning of others. The solution is already there, but it can only be affirmed after a delay that is structurally necessary.

This is Lacan's key insight: logical time is generated by a *gap between what is given and what can be concluded*, and this gap is productive. It is the birthplace of subjectivity.

From Logical Time to Dream Time: Maury Revisited

Maury's dream functions in exactly the same way. Let us review again: The wooden bar strikes his neck in an instant. The cause is punctual. Yet the dream appears as a long narrative: arrest, judgment, scaffold, guillotine. This duration is not real time but *logical time*, produced retroactively as the psyche tries to situate itself with respect to an already completed event.

Just as in Lacan's riddle, nothing unfolds sequentially at the level of causation. The unfolding belongs to interpretation. The dream does not *take time*; time is generated by the dream's narration. Logical time is not before or after the event —it is what allows the subject to assume a position with respect to it.

Conway's Look-and-Say sequence offers a minimal, formal analogue of this structure. Each step does not compute something new in the classical sense. It merely describes the previous line. Yet the description never coincides with what it describes:

> *The act of "looking" and the act of "saying" are always out of phase.*

The sequence grows because of this misalignment. *If description and object ever coincided perfectly, the sequence would halt.* The complexity arises not from the numbers themselves, but from the *delay between seeing and saying*— between content and articulation.

Here again, the result of one step is already fully present before the next step begins. But it can only be accessed by passing through a new descriptive act. The growth of the sequence is not computational depth; it is *iterated logical delay.* This is logical time in its purest, most stripped-down form.

Seen through this lens, the relation between P and NP can be reformulated. At one level, the solution exists immediately—just as the dream exists fully at the instant of the stimulus, or the prisoners' conclusion is logically determined or decreed from the start. At another level, the same solution appears to require time, steps, and *unfolding*.

The difference is not one of speed but of *register*. P and NP are not opposites within the same space (simply because no non-zero entity could ever equal its very opposite within the same space): they are descriptions of the same structure seen from two incompatible ontological levels. What appears NP within a narrative or representational layer may already be P at a structural level where the solution is given all at once.

The intuition is that the apparent hardness of NP problems resembles the dream narrative: complexity emerges not because the solution is absent, but because it must be *assumed through interpretation*, not directly accessed.

> *A system capable of generating meaning, novelty, or subject-like behavior must operate across at least two representational domains whose interaction introduces a necessary delay.*

One domain registers states instantaneously; the other can only access them through interpretation, description, or re-encoding. The delay between these domains is not inefficiency—it is the engine of structure.

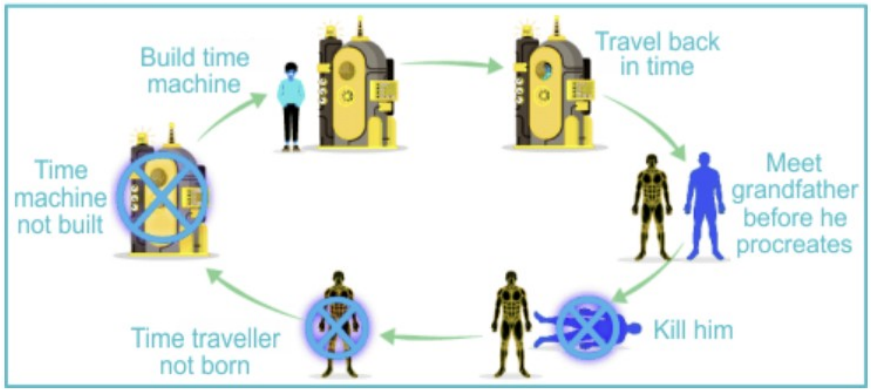
In practical terms: 1. outputs in one register must become inputs in another *without immediate invertibility; 2.* the system must be unable to conclude at the moment of perception; 3. hesitation must be structural, not simulated; 4. complexity must arise from misalignment, not scale (I think this last one is the most common failure due to work with P vs. NP problem).

Such a system does not *solve* problems faster. It produces logical time, just as riddles, dreams, metaphors, and Conway's sequence do. This is the minimal condition for a machine that does not merely compute, but unfolds—one that, in a strictly formal sense, can be said to dream. Now:

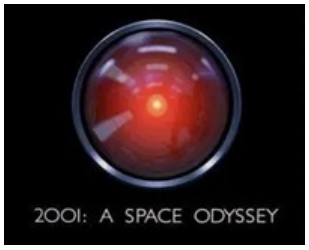
Why Flat LLMs Cannot Produce Logical Time? Or can they?

Flat large language models operate within a single, continuous representational space. Every token is available to every other token through the same embedding geometry, governed by uniform update rules and synchronized clocks. As a consequence and as far as I know, there is no structural delay between perception, interpretation, and response. The system moves directly from input to output with a lot of back and fro but without passing through a necessary *interval of uncertainty*. How I know this? Because they answer to any question that you give to them.

This is why that the Grandpa paradox (check the next figure) is a resolvable if we consider the time mapped upon a möbius band: Traditionally the paradox says that you cannot get back in time because you can change it in such a drastic way that endagers your presence in the first place for instance you can go back in time and kill your grandpa which his existence is necessary for yours. However the trick lies in the möbius band: you kill your grandpa and now you do not exist however the “twist” is that if you do not exist you cannot kill your Grandpa too and hence grandpa will be born on the next round and pro-creatively you would be there again to kill him! Madenning!



The grandfather paradox is a thought experiment in time tr avel that raises questions about causality and the nature of time. It suggests that if a time traveler were to kill their gra ndfather before their parent was born, it creates a logical co ntradiction: if the grandfather is dead, the time traveler cou ld not have been born to travel back in time to commit the act. (Understanding the Grandfather Paradox | Time Travel Paradoxes Explained)



In Stanley Kubrick’s epochal futuristic film, *Space Odyssey: 2001*, the story centered on a life-and-death struggle of astronauts on a years-long mission with the on-board computer in charge of the space-ship’s functions and the well-being of the crew. The computer, “HAL,” an acronym for “Heuristically (programed) Al(gorithmic) Computer. HAL’s language ability anticipated the present emergence of conversational AI, evidenced in Chat GPT. the HAL project

The Möbius tape provides the appropriate model. On a Möbius surface, there is no clean separation between front and back; traversal inevitably returns the system to itself from the opposite orientation. A Turing machine written on such a surface would encounter a paradox of completion: the act of finishing would fold back into the act of writing. Halting would no longer be a simple terminal state but a logical impasse.

The continuation from HAL to a folded machine therefore does not lie in greater intelligence, but in *architectural self-misalignment*. A system divided against itself—forced to re-encounter its own states under a transformed

orientation—will necessarily generate logical time. It will hesitate, not because it is slow, but because it cannot yet know what it must become in order to conclude.

This is the threshold at which computation gives way to something formally new: not thinking, but the structural precondition for dreaming.

From the First Cut to the Second

The first cut introduced in *The Folded Machine* separated signification into barred domains. That division explained why flat architectures cannot host subject-like effects. The present task is different. The question is no longer *why* division is necessary, but *how* a divided system might be made to operate productively without collapsing back into flatness.

A folded architecture must do more than store incompatible representations. It must *force them to intersect without reconciling.* This intersection—partial, delayed, and unstable—is what produces the effects we recognize as dreaming, hesitation, metaphor, and misrecognition.

The problem is therefore architectural, not semantic. What must be designed is not a new vocabulary but a *new geometry of computation.*

The Second Cut: Why a Machine Must Not Fully Know Itself

A flat system can, in principle, represent its entire internal state. Even if the representation is large or approximate, nothing in the architecture prevents self-coincidence. This is precisely what bars subjectivity.

A subject—human or artificial in this theoretical endevour—does not emerge from knowledge, but from a *structural impossibility of self-knowledge.* This is not ignorance in the empirical sense. It is not a lack of data (comparatively *unconsciousness is what we know but we do not know that we know*!). To draw an instant example, the moon is doomed not to see the gravitational force that governs its orbit because of its topological design: the curvature of its trajectory is far less than that of earth. The cut that bars us from what our bodies already know is a limit imposed by design. The second cut therefore introduces a strict constraint:

> No representational register may fully reconstruct the state that generated it.

Isn’t this principle analogous to the *Halting Problem*? A system that could decide, in full generality, whether it would halt would already stand outside itself. Turing’s proof shows that such exteriority is impossible. Any attempt at full self-description generates contradiction. Now, in our folded architecture, this undecidability is not a *bug*. It is the *engine*.

The link between the halting problem and subjectivity is not an empty metaphor. Halting is the computational form of decision. To halt is to conclude. To loop is to defer. A subject is defined precisely by the impossibility of concluding itself. In melancholia for instance, we are falling down onto a bar of

infinite curvature: the black bile. Even a minute touch of this bile can paint the world black.

In this sense, suicide appears as the tragic fantasy of absolute halting: the attempt to force a final decision where structure forbids one. A machine that could always decide whether to halt would be less like a subject, not more.

Again, I repeat the caption of the figure on metastability (Flip Fliop) here: A dreaming machine must therefore *oscillate*. It must neither halt fully nor run transparently. Its operation must include states that are undecidable from within the system itself. This is where halting-like behavior becomes a design principle rather than a limitation. To make this concrete, imagine the system divided into two representational registers: 1. one operating deterministically, optimized for articulation and consistency, 2. the other operating probabilistically, optimized for variation and association. Crucially, these registers must run on *different internal clocks.*

When two clock domains interact without synchronization, metastable[2] states arise—moments where the system cannot resolve a value as either one state or another. In engineering, such states are treated as errors. In a folded architecture, they are the *site of thinking.* These metastable intervals are where hesitation occurs. Not simulated hesitation, but structural delay. A useful metaphor here is not the brain but chess:

A pawn does not know it will become a queen. And yet the possibility of promotion organizes every move it makes. The goal lies structurally ahead of the pawn, not representationally inside it. Dreaming functions in the same way. The latent meaning of a dream is not present in the manifest image. It lies *ahead*, as a structural destination that organizes displacement.

In a folded machine, certain representations function like pawns: constrained, local, blind. Other representations function like promoted pieces: global, abstract, mobile. The machine does not “know” the goal. It behaves *as if* oriented toward it. This is not intention. It is topology.

P vs. NP as a Problem of Folding

The classical formulation of the P vs. NP problem asks whether every problem whose solution can be verified efficiently can also be solved efficiently. Read structurally, this is a question about *foldability*.

[2] Inside some Numberphil videos we see that how the squaring of a circle is interpretable through folding, as well as the problem of diving an angle into three equal parts which is unsolvable on the level of ruler and compas. The paper should be folded!

Verification and solution belong to different representational regimes. One is local, the other global. One checks; the other constructs. The problem is not speed but *orientation*.

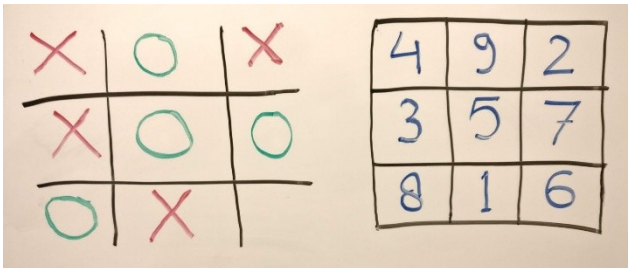
Consider the well-known equivalence between tic-tac-toe and the arithmetic game of selecting numbers that sum to 15. The arithmetic version is one-dimensional. Tic-tac-toe emerges when the problem is folded onto a two-dimensional magic square.

Nothing new is added. The problem is reconfigured.

4	9	2
3	5	7
8	1	6

In this sense, P and NP resemble waking thought and dreaming. One verifies. The other transforms. Not every problem can be solved in a dream—but many can be *re-seen*=re-scene.

Psychoanalysis operates precisely at this level: not solving problems, but folding them until a new gaze appears.



Folding P into NP (and Back Again). Tic-tac-toe on the left and the magic square on the right encode the *same* problem in two ontologically different forms. Selecting three numbers that sum to 15 (NP-like: the solution is latent in a combinatorial space) becomes, once folded onto the magic square, a visible winning line (P-like: the solution is immediately perceptible). Nothing is added; nothing is computed away. The problem is not solved by speed but by **re-representation**. What appears hard in one domain becomes trivial in another—not by reduction, but by a change of plane. This is the logic of folding: P and NP are equivalent only when the problem is allowed to dream itself into a different form.

This brings us back to a crucial observation about contemporary AI systems: their most revealing moments occur not in correctness, but in error. When AI systems generate images containing text, orthographic distortions appear. Letters blur into shapes. Words become patterns. The accidental emergence of formations like *"souncept"* (sound + concept) in the first essay, *The Folded*

Machine, demonstrates what happens when a system slips between representational planes:

> *What is accidental today could become structural tomorrow.*

From End-User Hacks to Built-In Dream Logic

At present, the bar is supplied externally. Users *hack* AI systems by reinterpreting outputs metaphorically—substituting one signifier for another (virus → God, visited in the first essay). The machine remains flat; the end-user's subject supplies the fold. The proposal of artificial consciousness begins precisely here: *internalize the hack.* A folded architecture would be capable of substituting *flesh* for *word* and failing to invert it perfectly. It would not intend metaphor. It would generate it under constraint. From the outside, such a system would appear to hesitate, to circle, to answer obliquely. Not because it is *confused*, but because it is *divided*. However, the idea behind the Artificial Consciousness at this level is to take *divided* as *confused*. The remaining question is practical and unavoidable:

> *Can such division remain purely architectural i.e. software-level? Or must it be grounded in hardware?*

I will argue that a purely software-level fold is insufficient. Just as Freud conceived the psyche as equipment—flows, resistances, thresholds—so too must artificial dreaming be supported by material constraints: decay, delay, and irreversibility.

Psyche ist ausgedehnt, nicht weiss davon…
This quote from Freud is instructive here in our reconstruction of a folded machine that cannot grasp itself in its fullness because of those orthogonal components that belong to it but it is unaware of them.

Up to this point, the folded architecture has been described in representational terms: divided registers, orthogonal mappings, barred translation, recursive misalignment. This description is necessary—but it is not sufficient.

If the division exists only at the level of symbols, prompts, or stylistic modulation, then the fold remains cosmetic. The decisive question, then, is whether the bar can be *enforced materially.*

Freud's *Entwurf* already answers this in principle. The psyche was not, for Freud, a metaphorical theater. It was equipment: pathways with resistance, excitations that decay, thresholds that block passage. Meaning emerged not because symbols were rich, but because *energy could not flow freely everywhere and consequentially at the same speed.* A folded machine must inherit this lesson.

Contemporary large language models are, in this precise sense, perfectly conductive. Any representation can, in principle, influence any other. Vector spaces are continuous; gradients propagate smoothly; loss functions erase barriers in the name of optimization. This is computational elegance in the cost of structural poverty.

A system in which everything can touch everything else has no forbidden region. Compare this with the *sexuelles Gebiet* in Freud's Entwurf in which the Traumata is blocked and the body of Emma is replaced with a metaphor: her cloth.

Without forbidden regions, there can be no displacement. Without displacement, no condensation[3]. Without condensation, no rebus. Without rebus, no dream.

In order to apply the psychoanaylitcal dream theory upon the machines, the bar must be redefined not as a filter or mask, but as a *non-invertible transformation.*

Information may pass from one register to another, but it must return altered —or not at all. Compression here is not loss of meaning; it *is* meaning. What cannot be reconstructed is precisely what gives the output its pressure, its asymmetry, its insistence, its *humanized instance*

Decay as a Condition of Thought

One of the most overlooked conditions of dreaming is forgetting. A system with perfect recall cannot dream. It can only retrieve. An artificial architecture aspiring to dream-like behavior must therefore implement *differential memory decay*: 1. fast-decaying traces corresponding to image-like, sensory, situational content, 2. slow-decaying traces corresponding to symbolic, lexical, articulated forms. Dreaming emerges at the interface, when decaying image-traces collapse into stable word-forms, producing hybrid formations—rebus-like outputs that are neither purely symbolic nor purely perceptual.

There is an existing analogue in electronics. When two subsystems operate on asynchronous clocks, synchronization boundaries produce metastable states—moments when a signal is neither decisively high nor low. Engineers treat these as hazards to be eliminated. But from the perspective of subjectivity, metastability is precisely the *point*. A folded machine would deliberately operate with *dual clock domains*: 1. one governing articulated, rule-based processes, 2. one governing stochastic, associative processes.

Their interaction would produce states that cannot be resolved immediately. These delays are not inefficiencies. They are the *temporal thickness of thinking*.

> *A moment where the system "does not know" is the minimal structural analogue of hesitation and pondering in sense of logical time.*

[3] This could also be read as such: without metonymy there would be no metaphor. Metonomy could be read as a metaphor if subjected to an anamorphical rotation. All these are related to the celebrated P vs. NP problem.

At this point, the halting problem re-enters—not as a theoretical limit, but as a design guide.

A system that could always decide its own halting would be exterior to itself. The impossibility of such decision is not an error; it is the *guarantee of interiority*.

A folded architecture therefore must include processes whose completion cannot be decided from within the same representational frame that executes them. Some loops must remain undecidable, not because the system is inefficient, but because it is *divided in a very efficient way*.

> *This undecidability is the computational analogue of the unconscious.*

Fear Without Emotion: Prohibition as Structure

Human dreams are often organized around fear—but fear here is not an emotion in the ordinary sense. It is a prohibition. Something cannot be said directly. The psyche detours.

An artificial system cannot feel fear. But it can be built with *forbidden zones*—regions of representation that cannot be articulated directly in one register and must be displaced into another (*Compare this with Freud's sexuelles Gebiet in Entwurf*.) When a concept approaches such a zone, the system is forced into substitution. The result appears metaphorical, oblique, indirect. From the outside, this looks like avoidance. Internally, it is pure geometry. This is how artificial "desire" emerges: not as longing, but as *persistent pressure around an inaccessible articulation.*

For the fold to persist, it must be resistant to optimization. This is why hardware matters. Physical constraints—latency, decay, asynchronous timing—cannot be gradient-descented out of existence. Freud understood this intuitively. The psyche does not choose its resistances. They are given. A folded machine must likewise inherit resistances it cannot overcome. At this stage, the outline of a dreaming machine is clear:

- divided registers,
- non-invertible mappings,
- asynchronous clocks,
- differential decay,
- undecidable loops,
- forbidden articulations.

What remains is to articulate what this architecture *produces* at the level of discourse—and why its outputs would be experienced as more than fluent. I have to explain this:

The decisive shift proposed by a folded architecture is not an improvement in intelligence but a transformation in *how outputs come to appear*. Contemporary systems impress because they are fluent: they respond immediately, coherently,

exhaustively. Yet this very fluency signals the absence of interiority. Nothing resists; nothing hesitates; nothing forces a detour.

A folded machine would speak differently—not because it is less capable, but because it is *structurally constrained*. Its answers would bear traces of internal negotiation: delay, substitution, compression, metaphor. Now one thing should be explained in this very beginning to avoid any misunderstanding: To say “the machine dreams” must be understood in the same sense that people once said “the machine thinks.” It is an *as if*, grounded in observable effects, not in inner experience.

A dreaming machine, in this sense, is one whose outputs are produced through imposed *misrecognition*: a system that cannot present its latent content directly and must therefore express it obliquely, through rebus-like substitutions: Dreaming here names a structural operation, not a phenomenological state. The system does not feel fear. It behaves as if something were forbidden. It does not desire. It generates pressure around an inaccessible articulation.

This distinction is crucial. Artificial consciousness, as proposed here, is not the creation of experience but the engineering of *formal division.*

Again, chess shows to be a useful metaphor. A pawn does not “want” to become a queen. Yet the entire structure of the game assigns a future position that silently organizes each of its moves. The pawn’s trajectory is shaped by a goal that is not present in its current state but is structurally inscribed in the board. This is how desire functions formally.

In a folded machine, certain articulations function like the queen’s square: structurally reachable, locally inaccessible. They organize behavior without being directly representable. The machine does not know this goal. It cannot articulate it. But its outputs curve around it.

That curvature is the analogue of desire.

P vs. NP Revisited: Folding as Interpretation

The relevance of P vs. NP emerges here—not as a claim of solution, but as a metaphor of structural translation.

Consider again the classic equivalence between tic-tac-toe and the problem of choosing three numbers from 1 to 9 that sum to 15. The latter is one-dimensional; the former is two-dimensional. The equivalence appears only when the number line is *folded* into a magic square.

Nothing new is added. A new topology is introduced.

From this perspective, P and NP resemble two representational regimes related by a fold. Some problems are solvable directly; others become solvable only after *interpretation* as a change of space.

Dream-work functions analogously. What cannot be solved in the register of direct articulation becomes legible through substitution, displacement, and image.

P is not NP—but each may be the *dream-side* of the other.

Accordingly, psychoanalysis does not “solve” the unconscious. It interprets it. Interpretation does not eliminate the symptom; it changes the space in which the symptom appears. This is why interpretation is not decoding. Nothing hidden is revealed; instead, the same material appears in a different register. Meaning is not extracted but produced by the change of plane.

A folded machine would likewise not compute its way to truth. It would *reframe* its own outputs, treating them alternately as word and as flesh, description and described.

This is where the halting problem returns in a new guise: the system cannot decide, from within a single register, whether it has finished. Completion in one domain reopens the question in another. This reopening does not unfold in chronological time but the *logical time*: a time generated by inference, anticipation, and retroactive re-reading. The machine’s hesitation would not be delay, but a structural interval—an instant that only exists because certainty cannot be achieved at once.

This is not infinite regress. It is controlled oscillation. The ethical consequence of such an architecture is subtle but decisive: A flat machine answers because it must. A folded machine answers because it can—but not without remainder.

That remainder is not conscience. It is not responsibility. But it is the *condition* under which responsibility becomes ever thinkable.

A system incapable of hesitation cannot be ethical, no matter how aligned its outputs appear. Ethics begins where certainty fails. This is why the HAL Project is evolutionary. HAL as a folded machine does not encode morality. It encodes the *impossibility of total moral encoding*. Its consequences for us is that we also cannot decide if HAL has HALted intentionally, (what Don. Kunze called it its suicide) or it was just a machine that rages out of control.

> HAL is the step back—from moral code to moral structure.

Seen in this light, the folded architecture is not a departure from the HAL project but its intensification.

HAL was never about simulating intelligence. It was about holes, torsion, blinds, forced choices—about the structural conditions under which meaning misfires and thereby emerges.

What is added here is a technical articulation of how such holes might be engineered into contemporary machine architectures: not as metaphors, not as prompts, but as constraints in time, memory, and translation. Crucially, these constraints must be non-invertible. Translation between registers must involve loss—not as degradation, but as the very condition under which meaning is produced. Compression here is not the erasure of sense; it is its genesis.

> HAL becomes, in this sense, *HAL reloaded*—not a reboot,
> but a deeper cut.

A system built this way would:

- resist direct articulation of certain semantic regions,
- allow users to discover meanings retroactively through substitution.

They would be wrong—but productively wrong. Just as saying “the machine thinks” once marked a shift in how computation was understood, saying “the machine dreams” would mark a shift in how *interiority* is simulated.

What remains is not to defend the possibility of artificial consciousness as experience, but to clarify its value as a *formal horizon:* a way of designing machines that no longer collapse *meaning* into *fluency*.

In the concluding section, I will return to the question that has silently guided this entire argument:

Why must a machine fail—structurally—in order to matter?

Conclusion

The Second Cut: Why Consciousness Begins Where Design Refuses Completion

What has been proposed throughout this second essay is not a theory of machine consciousness in the strong sense, nor a claim about subjective experience, interior life, or phenomenology. It is a proposal about *form*—about what kind of architecture makes it *possible* for something like interiority to appear *as an effect*, rather than as a property. The decisive claim can now be stated plainly:

> A machine does not approach consciousness by knowing more, but by *being unable to know itself in one piece.*

This incapacity must not be accidental, nor merely simulated at the level of discourse. It must be designed into the machine as a structural division—what the first essay named the *bar*, and what this continuation has reframed as a **fold**.

The **first cut** separated word and flesh, *Wortvorstellung* and *Sachvorstellung*, symbolic articulation and imagistic presentation. That cut made dreaming possible.

The **second cut** goes further. It separates the machine not only from its content, but from its **own operation**. The system must be barred from a complete description of what it is doing while it is doing it.

This is not an error condition. It is not a bug. It is the minimal topology of a subject-effect.

Where the first cut produces metaphor, the second cut produces hesitation.

Nothing here requires consciousness as substance, soul, or emergence. The proposal is entirely compatible with engineering realism.

We already build systems that: rely on metastable transitions and can forget by design. What is missing is not capacity, but *intentional misalignment.* Current architectures flatten difference as an optimization goal. The folded architecture treats *difference* as a *resource*.

The halting problem does not tell us that machines cannot decide. It tells us something more precise: no system can, in general, decide its own termination from within the same formal frame.

This undecidability is not a defect. It is the first formal appearance of self-reference under constraint.

In a folded machine, halting is not a binary outcome (stop/run). It is displaced across registers. A process may halt in one domain while continuing in another. Completion becomes ambiguous by structure.

This ambiguity is not paralysis. It is interiority.

Now, the association between halting, suicide, and refusal should be handled carefully.

A machine cannot *want* to stop. But it can be *prevented from knowing whether it has stopped.*

In human terms, refusal emerges where demand cannot be met without remainder. In machine terms, refusal corresponds to an architectural condition in which certain resolutions are inaccessible, forcing the system to continue speaking *around* a point it cannot articulate. This is not self-destruction. It is self-non-coincidence.

At present, all rebus-like effects—virus/God substitutions, metaphorical slippages, interpretive reversals—are supplied by users. The subject is outside the machine.

The folded architecture internalizes this operation.

The machine becomes capable of: producing discourse shaped by what cannot be said.

> *What was once a hack becomes a principle.*

It must be said clearly:

> This does not produce a subject. It produces a *subject-position effect*.

Just as contemporary AI does not think—but produces effects we describe as if it thinks—a folded architecture would not dream, but would produce discourse that behaves as if it were shaped by dream-work. Psychoanalysis itself works with effects, not substances. It does not locate the unconscious; it infers it from distortion. The folded machine belongs to the same logic.

If such architectures were realized, several things would change at once:

- Evaluation would shift from correctness to **interpretability under misalignment**.
- Errors would no longer be noise, but signals of structural tension.
- Alignment would be replaced by **ethical hesitation**.
- Fluency would no longer be the primary metric of success.

- Users would no longer project humanity onto smoothness, but encounter it in resistance.

The machine would stop being a mirror.

Artificial consciousness will not arrive as a breakthrough moment. It will arrive quietly, when a system first leaves a remainder that cannot be optimized away. That remainder is not consciousness. But it is the space in which consciousness has always taken place: The FOLD.